Filling the perceptuo-motor gap

Jean-Luc Schwartz

Filling the perceptuo-motor gap

Jean-Luc Schwartz

Institut de la Communication Parlée (ICP)
GIPSA-Lab Grenoble Image Parole Signal Automatique – Dpt Parole Cognition
CNRS UMR 5216, Université de Grenoble, France.

Introduction

The session under focus deals with phonetic articulatory-acoustic variation and with the link between phonetic detail and phonological modelling. The underlying question was to estimate the impact of phonetic detail on the phonological status of a given unit, and to attempt to better understand how details are produced by the speaker or exploited by the listener to access (encode or decode) the phonological level. But what is over and over at work in the four papers of this session is the possibility that there could exist a *gap* between the speaker's intention and the listener's perception, and that the phonetic variation is in some sense contained, produced or at least made possible by this gap. The present discussion will be focused on the perceptuo-motor gap. In an initial section, I will briefly recall how speech communication theories deal with the perceptuo-motor link. Then, I will discuss each of the four papers of the session, around a single question – *What does a listener know about a speaker's gesture?* – that is, what does the corresponding study tell us about the perceptuo-motor gap. Finally, I shall conclude around the theory that we have developed at ICP, called PACT (for "Perception-for-Action-Control Theory") in an attempt to show how it could indeed contribute to "*fill the perceptuo-motor gap*".

1. The perceptuo-motor gap in speech communication theories

The debate between auditory and motor theories of speech perception is ancient but still quite vivid. Auditory theories (e.g., Massaro, 1987; Nearey,1997) consider speech perception as a signal processing / pattern recognition problem, which should be considered in reference to the characteristics of the acoustic input, and the properties of the auditory system. The way signals are produced by the speech motor system is not considered as relevant for solving the task. On the contrary, motor theories (Liberman and Mattingly, 1985; Fowler and Rosenblum, 1991) assume that the listener recovers the speaker's gesture. The focus is not put here on auditory processing, but rather on this recovery mechanism, conceived in some sense as "integral", that is the phonetic percept *is* the speech gesture. Simplifying somehow, auditory theories, considering speech perception without action, posit a "no-link" between perception and action; while motor theories, considering speech perception without audition (this is most clearly expressed in the "speech is special" view, according to which audition does not intervene per se in the processing of speech gestures: see, e.g., Whalen and Liberman, 1987; Whalen et al., 2006) posit a "full-link" between perception and action. Hence in both kinds of theories, there is in fact *no gap* between perception and action in the speech communication process (Fig. 1).

Quite on the contrary, Ohala's theory of sound change sets the perceptuo-motor gap at the centre of language evolution: "Speakers exhibit variations in their pronunciation which they and listeners usually do not recognize as variation. When pronunciation is transmitted, however, the existence of this variation can create ambiguity and lead to the listener's misapprehension of the intended pronunciation norm. A misapprehended pronunciation is a changed pronunciation, i.e., sound change" (Ohala, 1989, pp. 175-176). Therefore, it is

precisely because the speaker's articulatory intention is not fully recovered by the listener that there is a sound change shaping human languages: the perceptuo-motor gap is central there. Lindblom's Dispersion Theory (Liljencrants & Lindblom, 1972) and its later Adaptive Variability Theory version (Lindblom, 1986) do not precisely deal with the perceptuo-motor link, but they rely on the same kind of initial principle: since the information on the acoustic-auditory stimulus may not be sufficient to recover the gesture, or rather the phonetic category of the uttered speech sequence, languages should select sufficiently different items to communicate. The perceptuo-motor gap is implicit in this framework.

Last but not least, Stevens's Quantal Theory (1972, 1989) can also be considered as a *theory of the perceptuo-motor gap*. In Stevens's view, the articulatory-to-acoustic/auditory relationship is nonlinear. This means precisely that the articulatory gesture cannot be exactly recovered by the listener, since a whole bunch of articulatory gestures provide quite the same sound. This nonlinear behaviour sets the place for uttering gestures that do not need to be finely tuned, which minimizes the need for a too high level of speech production accuracy. On the other side, the listener's task is also facilitated, since there are some kinds of "natural categories" of speech sounds. Hence, there is in the Quantal Theory a nonlinear behaviour setting the perceptuo-motor gap through which perception shapes action, and the substance of speech may shape the forms of language (Fig. 1).

We proposed another view of the perceptuo-motor gap, that we called the Perception-for-Action-Control Theory (PACT: Schwartz et al., 2002, 2007). This will be presented in the last section, in the hope that it could integrate some of the theoretical arguments presented above, together with some of the experimental facts discussed below.

2. What does a listener know about a speaker's gesture? Four case studies

I shall now take, one after the other, the four presentations of the session and systematically discuss the results in reference to the nature of the perceptuo-motor link and to the elements of answer they could provide to the question: "What does a listener know about a speaker's gesture?".

2.1. Solé & Ohala: What does a listener know about time vs. height control?

The basic assumption tested by Solé & Ohala is that the implementation of the height control in vowel systems is accompanied by durational differences that could then be taken into control by the speaker to enhance height contrasts. The consequence would be that low vowels, possibly realized as longer than high vowels because of the intrinsic longer duration of the opening gesture setting the adequate tongue height, would then be *controlled as longer vowels*. This acquired control would serve an auditory objective: enhancing the contrast between high-short and low-long vowels. In this process, the listener would reinterpret an intrinsic (biomechanical) consequence of a given speaker's command (a height value) as an explicit additional command (a duration value) and then use it as a speaker for controlling vowel production. Therefore, this clearly fits into the "sound drift from perceptuo-motor gap" scheme developed by Ohala, as recalled previously: duration, initially a biomechanical by-product, would be perceptually "misinterpreted" as an explicit control.

In respect to the basic claim in this target paper, that intrinsic vowel duration could be under the speaker's control in a language-specific way, two aspects of the data presented by Sole & Ohala are quite convincing. Firstly, since the relationship between height and duration seems language-dependent, different from American English and Catalan to Japanese, something has to change – and hence be controlled – from one language to the other. Secondly, the listeners seem to perceptually exploit duration in category identification. However, I would like to raise two questions about the Sole & Ohala's paper, one is theoretical and the other is methodological.

Perceptual integration of duration and height: question about a perceptual mismatch

Why (or how) could duration have been explicitly introduced as a control in vowel production? My first point is that perceptual misrecovery is not necessary there. A control may appear in the phonology of a language because it happens as a byproduct of speech production mechanisms (say for coarticulation reasons in general) and the speaker chooses to take control over this initially uncontrolled phenomenon. Is perceptual misrecovery needed in the case of vowel duration? I have an *a priori* concern with this assumption. The literature on vowel reduction suggests that height contrasts are seldom perfectly achieved, with contrasts between two height values often blurred by dynamical implementation and reduced articulation. On Fig. 2, I recall, just for the example, the data obtained by Loevenbruck & Perrier (1997) on the /ɛ/ vs. /a/ contrast, that is a mid-low vs. a low vowel in French. The study dealt with various conditions of focus and speed in a [iVi] context (Fig. 2a). Classically, the F1 value reached for the centre vowel V in unfocussed and/or quick utterances of [iai] were merged with those for focussed and/or slow utterances of [iɛi] (Fig. 2b). In this kind of control, it is very important that the listener is able to integrate current duration and current height (i.e. F1 achieved value) in order to be able to recover the target, even not reached.

Loevenbruck & Perrier (1997) showed that when the listener hears the whole trajectory, he/she is able to categorize the target properly why it is not the case if the centre of the trajectory with just the most open period is provided to the listener. Furthermore, they showed that the target recovery process seems compatible with a dynamic inversion mechanism, thanks to which a biomechanical model could enable to recover the target from the current trajectory, even in reduced cases. The lesson of this, in my view, is that it would be quite counterproductive for a listener to separate height from duration and recover control parameters from each of these variables independently. On the contrary, my guess is that time and height are integrated by the listener in a hopefully smart way. This integration process would involve knowledge of articulatory *dynamics*, rather than capitalize on purely *cinematic* variables.

Variation of duration with height: question about a "stable durational ratio"

Let me now address the experimental material *per se*. The study by Solé & Ohala concerns possible differences in duration between various height degrees as a function of speaking rate, for vowel contrasts in three different languages, i.e. Catalan, American English and Japanese. In all this study the target vowel (with variable height values) is produced from a closed configuration [b], within either an open syllable [bV] (Catalan and Japanese) or a closed syllable [bVd] (American English). Actually, the prediction by the authors is not completely clear. To quote their reasoning, "If vowel duration differences are the result of the larger distance the jaw and the tongue have to travel for low vowels vis-à-vis high vowels, such differences should be relatively constant across changes in speaking rate. " (Case 1), while "a difference in control strategy would presumably be reflected in enhancing durational differences at slower speaking rates, when vowels are longer, in order to maintain a constant perceptual distance (i.e., constant ratios) across rates" (Case 2). However, the precise

prediction is not very straightforward, since in the expansion of Case 1, the authors acknowledge that "If undershoot affects high and low vowels in the same degree, then absolute differences in vowel duration may be somewhat reduced in fast speech, or even eliminated if low vowels – which involve a greater displacement – are affected by undershoot to a larger degree than high vowels". They conclude "In sum, the view that vowel durational differences are due to biomechanical factors would predict an approximately constant vowel duration difference in slow and fast speech or slightly smaller differences at faster rates". Actually, the prediction that the differences should stay constant or at most slightly smaller at fast rates is not really demonstrated, the more so considering the statement in the conference version of the paper: "In sum, the view that vowel durational differences are due to biomechanical factors would predict an approximately constant vowel duration difference in slow and fast speech if the vowel target is maintained, but the difference would be smaller or eliminated at faster rates if articulatory undershoot was present".

Therefore, there is some potential contradiction in the argument, and I will try to make clearer what the prediction could be in the biomechanical hypothesis, that is Case 1. The question is: in a biomechanical model, what is really the prediction about possible differences in duration between low and high vowels as a function of speaking rate? To discuss this question, I will capitalize on a classical – and very simple – motor control model, the "second-order" agonist/antagonist model, and a classical theoretico-experimental study of jaw movements by Nelson (1983). The second-order model is recalled in Fig. 3a. In this model, the target results from equilibrium between an agonist and an antagonist muscle described as stiffness-driven springs in competition (Perrier et al., 1996; in the framework of the Equilibrium-Point hypothesis, originally proposed for the control of limb movements: see Feldman, 1986). This model can be easily described in mathematical terms as a second-order damped linear system,

producing typical trajectories from a closed to an open target Λ1 (high) or Λ2 (low) as shown on Fig. 3b. In these trajectories displaying e.g. jaw movements from the closed target to one or the other of the open targets, the "co-contraction" summing the stiffness of the agonist and antagonist muscles is kept constant, which results in identical durations for both movements. However, keeping co-contraction constant means applying different forces, hence different muscular effort, with a larger effort for target Λ2 than target Λ1. There comes the beautiful study by Nelson. In this work, the author applies theoretical considerations to compute in various versions of the second-order model the relationship between effort E, distance D and time T, that is the effort E to apply to achieve a certain distance D in a certain amount of time T.

The relationship between E, D and T for a given version of the model is displayed in Fig. 4a. It appears that if the effort is maintained constant for a movement towards either a close target Λ1 at distance D1 or a far target Λ2 at distance D2 (with D2 > D1), the amount of time necessary to reach the target is longer for the further target Λ2: T2 > T1. Of course, it could be argued that the effort is not kept constant in realistic biomechanical movements. The data obtained by Nelson for up and down strokes of a subject's jaw during normal speech provide a kind of "natural distribution" of the relationship between effort, distance and time. On this figure, I have plotted in a qualitative way a mean behaviour. Interestingly – and not surprisingly – this line providing a time/effort compromise for various distances to achieve shows that *movement duration increases with the distance*. Fig. 4a also displays three possible configurations on this line, e.g. a distance of 2 mm, 5 mm and 10 mm. A 10-mm distance could be, in a very qualitative way, attributed to a slow [a], and a 5-mm distance to a slow [i]. The 2-mm distance would provide, in the reasoning, an illustration of the reduced configuration for both targets (Fig. 4b). On this basis, the portrait of duration vs. rate values

for [a] vs. [i] does conform to one of the predictions by the authors for Case 1, reducing absolute differences in vowel duration in fast speech. But this is indistinguishable from the prediction in Case 2. Actually, the display in Fig. 4c is compatible with most data provided by Solé & Ohala. Notice that in this framework, articulatory trajectories towards [a] vs. [i], even reduced, could be adequately interpreted by listeners as /a/ vs. /i/ thanks to an articulatory dynamic inversion process, as in Loevenbruck et al. (1997).

Of course, this reasoning is very qualitative. The interest is to make clear that the predictions of a biomechanical model – here, a second-order model together with Nelson's predictions – can adequately conform to data. A key piece in the reasoning concerns the amount of reduction that could be expected from one vowel to the other. Quoting Solé & Ohala, is it plausible that "low vowels – which involve a greater displacement – are affected by undershoot to a larger degree than high vowels"? The answer is likely positive, and it is compatible with a view in which duration per se is not controlled, but rather the degree of reduction of the target (e.g. Lindblom, 1990), or the tuning of the target necessary to increase perceptual distinctiveness (Moon & Lindblom, 1994). In this reasoning, a proportional decrease in duration with rate, as displayed in Fig. 4b, would not indicate a control of duration, but a control of target and target reaching.

Conclusion

My claim is that the "duration misinterpretation" assumption is not necessary in the Solé & Ohala's paper: there is no strong evidence for a perceptuo-motor gap there. This is not to say that height recovery is a perfect process, with a "full link" between stimulus and gesture as in motor theories. Indeed, if reduction is too strong, the recovery may fail (see e.g. Pitermann, 2000). Moreover, if recovery were complete, automatic and effortless, there would be no need

for a speaker to hyperarticulate, while the control of articulatory strength is obviously part of the speaker's program (see Moon & Lindblom, 1994). Hence comes the need to incorporate auditory processing in the recovery of articulatory targets, in addition to possible articulatory knowledge. As a matter of fact, the organization of height degrees and more generally the organization of vowel systems in human languages seems to clearly obey to listener-oriented principles, i.e. dispersion – that is distances between vowels in a formant space (Lindblom, 1986) – and focalization – that is grouping of consecutive formants for a given sound (Schwartz et al., 2005) – as displayed in our predictions of vowel systems in the Dispersion-Focalization Theory (Schwartz et al., 1997). In this framework, it is not surprising that duration can be explicitly controlled by speakers to enhance perceptual contrasts, which is the case in a number of languages, as recalled by Solé & Ohala. In conclusion, in spite of my general agreement with Solé & Ohala that vowel duration may be linguistically controlled, my own view of their data in terms of perceptuo-motor gap can be formulated in two points: (1) The listener is able to process current height and duration in order to separately recover the intended height control (vowel reduction)[1] and the intended vowel duration; (2) Recovery is constrained by both *articulatory knowledge* and *auditory processing*, as shown by the organization of sound systems in human languages.

[1] Solé & Ohala propose that "Schwartz, in the discussion to this paper, argued that the rescaling of the differences in duration between high and low vowels may be obtained with intrinsic timing models by integrating durational and spectral information in the target movements, such that differences in articulator's velocity for high and low vowels would be present. But of course, if duration is integrated in the specification of the vowel, it is extrinsically determined, as is precisely claimed in this paper". My point is not about extrinsic vs. intrinsic vowel duration. My basic claim is that even if duration is NOT integrated in the *specification* of the vowel, the listener is able to *integrate ON-LINE spectral and temporal information about the acoustic trajectory* to estimate the spatial target in spite of vowel reduction phenomena.

2.2. Kuzla, Ernestus & Mitterer: What does a listener know about voicing assimilation?

The study by Kuzla et al. capitalizes on a previous study of voicing assimilation processes in speech production. In that study, Kuzla et al. (submitted) explored the role of prosodic strengthening on voicing assimilation, that is devoicing of a word-initial voiced fricative /v/ or /z/ preceded by a word-final voiceless obstruent. They showed that there was more devoicing after prosodic word boundary than after phrase boundary. This was due, in their interpretation, to "domain initial strengthening" thanks to which segments at the beginning of higher prosodic domains are articulated more strongly than at the beginning of lower prosodic domains. Indeed, since a prosodic phrase boundary is a higher prosodic domain than a word boundary, it would lead to stronger articulation after the boundary, hence less assimilation and thus less devoicing. The goal of the study by Kuzla et al. in the present session was to determine if subjects were able as listeners to exploit the knowledge they had as speakers, and hence to compensate for devoicing processes in a way accounting for the prosodic strengthening effect. They showed that it seemed to be actually the case. This would suggest that, for voicing assimilation at least, there is no big gap between perception and action.

Integrating procedural knowledge about action in perception

Before entering into the data themselves, it is interesting to stress that the perceptuo-motor link discussed here is very different from the link discussed in the previous study. Indeed, while the focus was put by Solé & Ohala on the correct or wrong recovery of *biomechanical* effects, the work by Kuzla et al. deals with coarticulation facts, and more generally with *control mechanisms* that would or would not be incorporated by the listener in the speech interpretation process. This is reminiscent of a famous – but perhaps not so well known by

speech scientists – series of works by Paulo Viviani and colleagues on the perception of handwriting gestures.

The claim by Viviani & Stucchi (1992) is that the perceiver of a given human gesture is able to exploit "procedural knowledge about the repertoire of potential gestures" in the perceptual processing of the perceived gestures. To show this, the authors search whether biological motion is "special", and argue that it is, in the case of two-dimensional hand movements, characterized by what they call the "Law of Motion". This law describes the relationship that would exist between V(t), the instantaneous tangential velocity of the movement, and R(t), the corresponding radius of curvature of the trajectory. The law specifies that velocity increases with the 1/3 power of the radius: straighter portions are realized at a highest speed than more curved ones. Interestingly, this law of human gestures seems to be incorporated in human perception, in the estimation of both movement and shape. Firstly, to be perceived as uniformly quick, a movement has to respect the 1/3-power law: inside a hand gesture realizing an ellipsis, the speed must not be constant (as in Fig. 5a) to be *perceived* as constant. Even more puzzling, if a circle is traced according to a velocity profile mimicking the one from an ellipsis (Fig. 5b), the shape is perceived as elliptic rather than circular (Fig. 5c). This is the kind of "procedural knowledge" about motor control processes that could be exploited by the listener in the Kuzla et al.'s study (see also the various proposals by Fowler in the same vein: e.g. Fowler, 1986).

Discarding a pure auditory interpretation

To assess the perceptual data presented by Kuzla et al., I would like to play the role of the devil's advocate asking a simple question: is it possible that these data have a purely auditory origin, independent of any knowledge about speech motor control? Actually, there is an

auditory mechanism able to play a role in the present data, which is auditory forward masking. This is the effect produced by a given acoustic stimulus decreasing the audibility of a stimulus posterior in time, provided that the masking and the masked stimuli occupy similar regions in the auditory spectrum. Fig. 6a displays in a very schematic way some basic time-frequency ingredients of the stimuli studied by Kuzla and colleagues. I have focused on the crucial piece in the reasoning, which is the amount of voicing, represented in the figure as a voicing bar. In the case of a left /ə/-context, referred by the authors as the non-assimilation context, there is a voicing bar in the contextual portion, able to partly mask the voicing bar of the voiced obstruent under study, that is the word-initial /v/ or /z/ (Fig. 6a, left). On, the contrary, in the assimilation context with a voiceless obstruent before the word-initial voiced fricative, there is much less forward masking by the left-context voiceless obstruent /t/, since it does not contain energy in the adequate frequency region (see e.g. Moore, 2003, for a review of forward masking and its time-frequency characteristics). Therefore, forward masking provides a likely explanation for the "assimilation compensation" results obtained by Kuzla et al.: instead of assimilation compensation, the data would just reveal the differential effect of forward masking in the two conditions.

There are however some problems with this interpretation. A first one is that in some extreme cases (word-boundary assimilation condition in Experiment 2), listeners may judge a completely devoiced [v] as voiced in 20% of the cases: masking would not be of great help there, and Kuzla et al. rather consider the voicing of the previous [t] as a likely candidate for the explanation of this puzzling fact. More importantly, let us consider the forward masking predictions relative to the word- vs. phrase-boundary context in Experiment 2 (Fig. 6b). In this case, it is quite likely that the stronger articulation in the phrase-boundary condition should lead to both a longer pause between the left context (either /ə/ or /t/) and the word-

initial /v/, and possibly a rise in the initial F0. Both these factors should lead to *less masking* in the phrase condition, resulting in an increase in the audibility of the voicing bar, and hence a *higher* score of "voiced" responses. On the contrary, the articulatory procedural knowledge" assumption should lead to the inverse prediction: the phrase-boundary condition producing less assimilation, it should lead to *less compensation* in the listener's processing, and hence to a *lower* score of "voiced" responses. This is actually the case in the experimental data obtained by Kuzla et al. in their Experiment 2.

In summary, forward masking could explain compensation for assimilation, but not the specific role of prosody in this process. Other mechanisms related to Auditory Scene Analysis and involving continuities vs. discontinuities in spectral trajectories (e.g. Bregman, 1990) would lead to essentially the same reasoning that with forward masking. Notice that this does not discard the possible role of such auditory mechanisms in the first case: we know, at least since Repp's study on trading relationship in phonetic categorization, that both auditory and phonetic mechanisms may intervene in the same perceptual phenomenon, though differently from one configuration to the other (Repp, 1983).

Conclusion

In conclusion, the assumption proposed by Kuzla et al. gets reinforced by the devil's advocate reasoning I have attempted here, and my conclusion is similar to theirs. I would formulate it, in reference to the work by Viviani and colleagues, by the proposal that the listener exploits procedural knowledge on speech motor control in speech perception, at least for the processing of voicing assimilation phenomena (but also for many others, as recalled in their introduction). The complete study by Kuzla et al. shows that this procedural knowledge may be quite complex and precise, as displayed by the results of their Experiment 3.

2.3. Mielke, Baker & Archangeli: What does a listener know about the many-to-one sound-to-control recovery problem?

In the two previous studies, the question asked was to know if the listener was able to recover the speaker's commands from the sounds emitted by the vocal tract. Mielke et al. raise a serious problem in this framework: in some cases, the solution to the recovery problem is not unique hence it is illusory to expect a clear-cut solution, at least in computational terms. Actually, their study shows that in the case of the production of /ɹ/, various solutions are exploited by speakers with a large inter-individual variability (i.e., idiosyncrasy). This is obviously a quite large perceptuo-motor gap.

Speech robotics and the many-to-one articulatory-to-acoustic inversion problem

This relates to a classical problem in articulatory-to-acoustic inversion that is the fact that it is a many-to-one problem, with different articulatory configurations for a same produced sound (see e.g. Atal et al., 1978; Boë et al, 1992). The problem of recovering a command from the set of signals it generates is also a classical problem in robotics and it is at the core of a research program we contributed to develop in the 90s, that is speech robotics (Abry & Badin, 1996; Laboissière et al., 1991). Speech robotics proposes to apply the tools of cognitive robotics to the speech communication problem, acknowledging the fact that speech production consists in driving a set of actuators to produce a set of acoustic and possibly optic outputs, exactly as a robot has to drive a set of motor commands to solve a task specified by the values captured by its sensors (e.g. follow a wall, search for a light, track an intruder, capture an object, etc).

In speech robotics, the many-to-one problem is generally solved by introducing dynamic constraints, selecting a given command in a given context, to decrease the articulatory effort by smoothing the trajectory (that is, modelling coarticulation in a dynamic framework: see e.g. Bailly et al., 1991; Guenther, 1995; Guenther & Perkell, 2004). This is the kind of reasoning implied in Mielke et al.'s Experiment 2. However their Experiment 1 shows that the under-specification of the command from the sound may generate a true choice for the speakers, leading them to select different configurations for the same /ɹ/ acoustic output.

The developmental trajectory shaping articulatory-to-acoustic inversion: a case study

Apart from dynamics, there is a component of behaviour that could play a part in the selection of one command rather than another (the "regularization" problem), though probably not in the /ɹ/ case. This component, very seldom mentioned to my knowledge, is the *developmental trajectory*. I will illustrate this point, not for /ɹ/ on which I have no pertinent data – sorry for abandoning the field for a short while! – but on /u/.

Theoretically, human vocal tract is able to produce three types of [u] with identical first three formant values (Boë et al., 2000): velo-palatal, velo-pharyngeal and pharyngeal, whose main intra-oral constrictions are palatal, in the upper pharynx and pharyngeal, respectively. This means that if human languages select their vowel systems on the basis of acoustic dispersion (Liljencrants & Lindblom, 1972; Lindblom 1986), they could choose among various tongue configurations to achieve the corner vowel [u], which belongs to almost 95% of them. But, in nature, native (adult) speakers of all the tested languages produce the velo-palatal [u] only (Wood, 1979). The pharyngeal [u] has never been recorded. The velo-pharyngeal [u] has only been observed in a few speakers asked to utter [u] with a tube between their lips that compelled the participants to acoustically compensate for the lips opening (lip-tube paradigm,

Savariaux et al., 1995). This pattern has brought Abry and Badin (1996) to propose that the palatal [u] could be the first [u] production strategy explored during speech development. In other words, the displayed preference for the palatal [u] in adulthood would stem from its early sensori-motor mapping which would prevent the lip-tubed speakers from being able "to really abandon their acquired link between the [u] acoustic target and [the velo-palatal] position [of the tongue]" (Abry & Badin, 1996).

In a recent modelling study (Serkhane et al., 2007), we attempted to relate available acoustic data characteristic of infant vocalizations at 4 and 7 months with the behaviour of an articulatory-acoustic model of speech production incorporating vocal tract growth, the *Variable Linear Articulatory Model* (hereafter VLAM, Boë et al., 2002). Exploring acoustic vocalizations produced by 4-months-old infants enabled us to characterize the likely commands and vocal tract shapes at this age, and to explore the shapes corresponding to configurations in the [u] region. We compared the vocal tract shapes associated to the formants of an actual four-month-old vocalization in the [u] region, produced by the 4-month old VLAM, whether its whole set of parameters was used for production or only the subset of articulatory parameters and the range of articulatory commands compatible with actual vocalizations at four months was exploited. It appears that palatal to pharyngeal tongue highest points were displayed in the complete model while *only palatal* ones were displayed in the developmentally restricted model. This suggests that the first [u] around 4-month old would be palatal like the adult one: it is likely to be mapped early in life, thereby providing a precursor for this vowel in mature speech, which confirms Abry and Badin's hypothesis.

Conclusion

Coming back to /ɹ/, the lesson of the Mielke et al.'s study is quite important for a better understanding of what is or should be a perceptuo-motor link. Their data show *that a listener does not know everything about the speaker's gesture*: there is certainly a gap there. This suggests that gestures cannot be adequately defined without their perceptual counterpart. What makes an /ɹ/ an /ɹ/, after Mielke et al.'s data is not the vocal tract shape per se – nor the sound alone, recalling the role of procedural knowledge on action in perception – but the *perceptuo-motor coordination*, possibly with various motor configurations for a same sound. This coordination is learnt in the course of development, possibly in an idiosyncratic way, as in Mielke et al.; or possibly with a shape constrained by the developmental trajectory, as for /u/ in our own modelling work. A gestural theory should, in this context, incorporate a perceptual component – as could actually well be the case in the "global character of phonetic gestures" discussed by Mattingly (1990), already incorporating aerodynamic factors, in addition to vocal tract configurations conceived as global coordinations oriented towards a functional phonetic goal.

2.4. Gafos, Hoole, Roon & Zeroual: What does a listener know about the Place Order Effect?

The contribution by Gafos et al. deals with the role of time in gesture control in relation with phonology and grammar. It is not really focussed on the speaker-listener interaction – but rather on the possible intervention of grammar in the articulatory control of time in speech production. There is however one part of their study which does really involve the interaction between gestures and percepts, that is the study of the "Place-Order Effect" (POE). In that

piece of the studied data, Gafos et al. test the so-called place order hypothesis, according to which in consonantal clusters, if the second consonant is articulated at a more anterior place than the first one (e.g. “tp”) it cannot be anticipated too much so as to let the release of the first consonant audible. On the contrary, if the second consonant is articulated at a more posterior place (e.g. “pt”), even if it is anticipated, it lets the anterior part of the vocal tract open, and hence the first consonantal release is audible. This would result in more clustered sequences from front to back (e.g. “pt”) than from back to front (e.g. “tp”), which provides the place order effect. Though evidence for this is available in various studies, the results by Gafos et al. are mitigated. Indeed, the effect seems to depend on the speaker, possibly driven by inter-individual differences in the ability to exhibit more or less overlap between consonantal gestures within clusters: speakers or utterances allowing for more clustering would show a larger effect. This is what Gafos et al. propose to call the “relativized place order hypothesis”. Since the effect would be speaker- and utterance-dependent, they suggest that it is non grammatical.

The Labial-Coronal effect as a phonologization of the POE: an articulatory and acoustic study

I would like to attempt to relativize – and possibly re-grammaticalize or at least phonologize – this relativized non-grammatical version of the POE, capitalizing on a recent bunch of studies we have done about the Labial-Coronal (LC) effect. This effect refers to the significant trend in human languages to display more often in CVCV initial sequences a labial-to-coronal association (LC) rather than a reverse coronal-to-labial (CL) one, e.g., “pata” is displayed more often than “tapa” (MacNeilage & Davis, 2000). Interestingly, the LC effect is also found in infants' first words in the course of ontogeny. In the framework of their “Frame-Content Theory”, MacNeilage and Davis (2000) propose that the LC effect might result from the

conjunction of two principles in the course of speech development: (1) labial proto-consonants (or closants in the Frame-Content terminology) would be more simple to produce by infants than coronal ones, and (2) the production of their first words would lead infants to select a "simple first" sequence of actions beginning by the simpler gesture, supposed to be the labial one. The preference for LC sequences in development would then have been preserved in adults' languages. The work on articulatoy models at ICP with various vocal tract models differing in their morphology, lead us to discard the MacNeilage and Davis "simple-first" interpretation. Actually, there is no strong reason to believe that labial proto-consonants, considered as "pure frames" in the Frame-Content theory (supposedly uttered with no active articulatory command except jaw raising), are simpler than coronal ones, considered as "fronted frames" with an active tongue fronting gesture. Indeed, depending on the vocal tract model morphology, raising the jaw might produce a labial as well as a coronal contact (Vilain et al., 1999), and analyses of babbling inventories actually do not display a preference for labial patterns over coronal ones (Locke, 1983).

Our proposal is in fact similar to the one invoked by Gafos et al. for the Place Order Effect. The claim is that articulating a coronal after a labial would allow more anticipation than the inverse, due to asymmetries in coarticulation mechanisms, just as "front-to-back" clusters would enable more overlap than "back-to-front" ones. This would make LC sequences more simple and economic in articulatory terms than CL ones. To confirm this hypothesis, we exploited the speeding paradigm introduced by Kelso et al. (1986) to let more stable motor forms emerge from a competition between various possible behaviours. Therefore we investigated the coordination between the jaw, the tongue tip and the lower lip during the repetition with rate increase of Labial-to-Coronal CVCV disyllables (e.g. /pata/) and Coronal-to-Labial ones (e.g. /tapa/) by 28 native French speakers (Rochet-Capellan & Schwartz,

2007). It appeared that for the two types of disyllables the speeding process induced a shift from two jaw cycles per disyllable to one in a way that tended to limit the jaw energy consumption. Moreover, this shift modified the coordination between the jaw and the constrictors and came with a progression towards a LC attractor (e.g. /pata/-/tapa/ → /patá/ → /ptá/) or a CL attractor, (e.g. /pata/-/tapa/ → /tapá/ → /tpá/). Yet, the LC attractor was very significantly favoured among speakers regardless of the initial sequencing. Thus, the obtained results displayed a *strong statistical trend towards LC sequences*, though not a universal pattern, which is in line with the observation by Gafos et al. that the POE could vary from one speaker to another, though the prediction associated to the relativized POE linking the effect with speaker's overlapping preferences was not studied in our own work. This provides an interesting expansion of the Place Order effect to CVCV sequences.

The Labial-Coronal effect emerging in perception: a Verbal Transformation experiment

Since there seems to be an articulatory synergy chunking LC sequences rather than reverse CL ones within a single jaw cycle, could it be the case that the listener knows this speech production coordination and exploits it for perceptual chunking? This was the purpose of an experiment we did, using the Verbal Transformation Effect (Sato et al., 2007). This paradigm refers to the perceptual changes experienced while listening to a speech form cycled in rapid and continuous repetition, e.g. "life life life" switching towards the perception of "fly fly fly" (Warren, 1961). In our study, we examined the perceptual stabilities of repeatedly presented disyllabic sequences, involving either a labial-vowel-coronal-vowel (LC) or a coronal-vowel-labial-vowel (CL) phonological structure. Such structures can lead, in the Verbal Transformation paradigm, to a number of switches from e.g. the perception of /pata/ to the perception of /tapa/ or vice-versa. In two experiments on French listeners, involving either voiced or unvoiced plosive consonants, a greater stability and attractiveness was observed for

LC stimuli. Interestingly, while the French language is characterized by a preference for LC sequences with unvoiced stimuli, the pattern is reversed with a preference for CL sequences with voiced sequences, which discards simple phonotactic or lexical explanations of the observed asymmetry. Therefore, it seems that the larger stability and coherence of LC sequences, in line with the Place Order Effect for consonantal clusters, results in the fact that in a (…)CLCLCLC(…) flow the listener could more naturally provide a segmentation into LC chunks.

Conclusion

The results about the LC effect provide three elements of discussion related to the paper by Gafos et al. Firstly they back up the Place Order Effect, extending it to CVCV sequences, in a way compatible with the "relativized" version proposed by the authors, that is playing the role of a general trend with possible idiosyncratic variations rather than a systematic rule. Secondly, they suggest a possible link between the POE, considered as a non-grammatical fact by Gafos et al., and a trend in the sound systems for human languages, that is the preference for LC sequences over reverse CL ones. Thirdly and most importantly in the context of the main theme of the present discussion, they show that the listener incorporates the speaker's trend to chunk LC sequences inside perceptual processes, by perceptually chunking LC sequences rather than the inverse: perception and action seem closely coupled there.

3. Filling the perceptuo-motor gap in the Perception-for-Action-Control Theory

We are now left with a situation which might appear fuzzy. Indeed, while Sole & Ohala do not, in my view, really provide a demonstration that there is a perceptuo-motor gap in height-duration recovery for vowel perception, it appears that the listener seems to know quite a lot about voicing assimilation (in Kuzla et al.) and about Labial-Coronal chunking (in our expansion of Gafos et al.' Place Order Effect) for speech perception, just as Viviani & Stucchi suggest that perception involves "procedural knowledge" about the laws of action. However, the many-to-one relationship between gestures and percepts, recalled and displayed by Mielke et al., clearly discards perception as a simple mirror of action. This provides a picture at odds with both pure auditory and pure motor theories: there is a link between perception and action, but the link maintains perception and action different, which seems to result in … a perceptuo-motor gap!

The Perception-for-Action-Control Theory (PACT)

I employed the phrase " perception as a mirror of action" on purpose, in reference to recent developments about the mirror neurons observed in the monkey and possibly the human brain. Mirror neurons, displayed in the premotor area F5 of monkey brains supposed to be an homologous region of Broca's area in humans, fire in the same way when the monkey performs and sees an action (Gallese et al., 1996) or hears the sound of it (Kholer et al., 2002). A number of recent fMRI or TMS experiments confirm the role of motor or premotor areas in speech perception in humans (e.g. Fadiga et al. 2002, Watkins et al. 2003). This has been taken as providing experimental evidence in favour of the Motor Theory. For sure, it shows that there is a cortical and functional link between perception and action, which discards pure auditory theories with their "no link" characteristic. However, this does not

mean that perception is a simple mirror of action. Actually, a number of data show that *perception shapes action*. This is typically displayed in cases well captured by the Quantal Theory, where various types of gestures produce essentially the same sound, while a small modification of the gesture dramatically changes the sound. Just to take an example: slowly decreasing the lip area from an unrounded [i] first does not change the sound almost at all, and then suddenly dramatically changes it into an [y]-like sound, because of both acoustic and auditory reasons (Abry et al., 1989). Thus, "lip rounding" is not a purely "motor" concept, but rather an auditory-motor, and in fact, rather an "auditory-visual-motor" one.

The *"Perception-for-Action-Control" Theory* (PACT, Schwartz et al., 2002, 2007) proposes that the objects of speech perception are neither auditory nor motor but perceptuo-motor. Perceptuo-motor coordinations are learnt in the course of speech development, and they enable perception to control action – as is displayed in vocal imitation, appearing since four months of age (Kuhl & Meltzoff, 1996), or even earlier (Kugiumutzakis, 1999). Perception, auditory and visual, enables to learn and recover the appropriate controls for action. In this context, perception is shaped by the structure of action. But perceptual processes are not "transparent": they shape action in return, as in the rounding example provided previously. Perception in this sense also controls the actions of somebody else, by providing templates and prescriptions for the suitable actions of a given communication system. In summary, perceptuo-motor coordinations in PACT are both multi-sensory percepts regularized by the knowledge of speech gestures, or speech gestures shaped by perceptual processes.

A computational and cortical PACT circuit for filling the perceptuo-motor gap

The perceptuo-motor link is central in the PACT. Speech perception exploits the structure of action through this link, which agrees well with our discussion of Kuzla et al. or Gafos et al.

This does not mean that phonetic perception cannot occur before the perceptuo-motor link takes place (as in a number of early perceptual behaviour appearing before any speech production ability). It does not mean either that the perceptuo-motor link is always necessary: the involvement of motor areas in speech perception seems to depend on the difficulty of the task (e.g. Callan et al., 2003). The perceptuo-motor link could take the form of an analysis-by-synthesis model, as in Stevens & Halle, 1967; Stevens, 2002; or of a coordination of sensory and motor maps, as in Guenther's DIVA model (1995). We use a Bayesian probabilistic framework in which a distribution p(P,M) relating perceptual and motor parameters is learnt in the course of development, and then used for imitation and tuning to the ambient language (Serkhane et al., 2005).

A plausible cortical circuit for PACT is provided by the "dorsal route" (Hickok and Poeppel 2000), connecting perceptual multisensory processes in the temporal region (Superior Temporal Sulcus) with action understanding in the frontal lobe (including Broca's area and motor and premotor areas) passing by parietal regions matching sensory and motor representations (see, e.g., Sato et al. 2004). The important point here is that there is a possible cortical and computational framework for *linking* sounds and gestures in a functional way. This is the way the gap may be filled. Sounds and gestures are different objects, but they are related, cortically and computationally, in a principled and usable way for both producing and perceiving speech.

Conclusion

The focus I selected for the present discussion enables to reinforce a view that could be considered, after all, as rather classical. Indeed, as in many previous debates on the theme, it appears that relating various experimental data in terms of auditory vs. motor theories just shows that not everything can be accounted for by each of them. This is the reason why I consider a perceptuo-motor theory of speech perception, as the PACT, as just inescapable. Speech perception involves motor knowledge, without being pure motor recovery … just as speech production involves producing gestures, but gestures shaped by auditory (and visual) processes. A perceptuo-motor link structuring both speech perception and speech production is central there, "filling" the perceptuo-motor gap in a principled way. Filling the gap does not mean that the gap disappears. Actually, the perceptuo-motor gap is essential for both speech perception (including nonlinear pattern providing natural categories, as in the Quantal Theory of Speech Perception) and speech production (providing the basis for coarticulation, that is preparing gestures without audible consequences). The point is that speech perception and speech production need a perceptuo-motor link, just as any robotic system does.

However, whatever the reader will think about this theoretical reading of the present works, let me finally express many thanks to the contributors of this session for their very nice and rich set of data: after all, data sometimes pre-exist and always survive to theories!

References

Abry, C., & Badin, P. (1996). Speech Mapping as a framework for an integrated approach to the sensori-motor foundations of language. *4th Speech Production Seminar, 1st ESCA Tutorial and Research Workshop on Speech Production Modeling: from control strategies to acoustics*, 175-184, May 21-24, 1996, Autrans, France.

Abry, C., Boë, L.J., & Schwartz, J.L. (1989). Plateaus, catastrophes and the structuring of vowel systems. *J. Phonetics*, 17, 47-54.

Atal, B. S., Chang, J. J., Mathews, M. V. & Tukey J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *J. Acoust. Soc. Am.*, 63, 1535-1555.

Bailly, G., Laboissière, R., & Schwartz, J.L. (1991). Formant trajectories as audible gestures : an alternative for speech synthesis. *J. Phonetics*, 19, 9-23.

Boë, L.J., Abry, C., Beautemps, D., Schwartz, J.L., & Laboissière, R. (2000). Les sosies vocaliques. Inversion et focalisation. *XXIIIèmes Journées d'Étude sur la Parole*, Aussois, 257-260.

Boë, L.J., Heim, J.L., Honda, K., & Maeda, S. (2002). The potential Neandertal vowel space was as large as that of modern humans. *J. Phonetics*, 30, 465–484

Boë, L.J., Perrier, P., & Bailly, G. (1992). The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *J. Phonetics*, 20, 27-38.

Bregman, A.S. (1990). *Auditory scene analysis*, Cambridge MA: MIT Press.

Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, 14, 2213-2217.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience,* 15, 399–402.

Feldman, A.G. (1986) Once more on the equilibrium point hypothesis (λ model) for motor control. *Journal of Motor Behavior*, 18, 17-54.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *J. Phonetics*, 14, 3–28.

Fowler, C.A., & Rosenblum, L.D. (1991). The perception of phonetic gestures. In I.G. Mattingly & M. Studdert-Kennedy (eds.), *Modularity and the motor theory of speech perception* (pp. 33–59). Hillsdale, NJ: Erlbaum.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996) Action recognition in the premotor cortex. *Brain*, 119, 593-609.

Guenther, F.H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102, pp. 594-621.

Guenther, F.H., & Perkell, J.S. (2004). A neural model of speech production and its application to studies of the role of auditory feedback in speech. In: B. Maassen, R. Kent, H. Peters, P. Van Lieshout, & W. Hulstijn (Eds.) *Speech Motor Control in Normal and Disordered Speech* (pp. 29-49). Oxford: Oxford University Press.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Science,* 4, 131–138.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *J. Phonetics*,14, 29– 59.

Kohler, E., Keysers, C., Umiltà, M.A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science,* 297, 846-8.

Kugiumutzakis, G. (1999). Genesis and development of early infant mimesis to facial and vocal models. In J. Nadel et G. Butterworth (Eds), *Imitation in Infancy* (pp. 36-59). Cambridge University Press.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: vocal imitation and developmental change. *J. Acoust. Soc. Am.*, 100, 2425- 2438.

Kuzla, C.,, Cho, T., & Ernestus, M. (submitted). Prosodic strengthening of German fricatives in duration and assimilatory devoicing. Max Planck Institute for Psycholinguistics, Nijmegen.

Laboissière, R., Schwartz, J.L., & Bailly, G. (1991). Motor control for speech skills: a connectionist approach. In D.S. Touretzky, J.L. Elman, T.J. Sejnowski, & G.E. Hinton (eds.) *Connectionist Models, Proceedings of the 1990 Summer School* (pp. 319-327). San Mateo CA : Morgan Kaufmann Publishers.

Liberman, A.M, & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition,* 21, 1–36.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language,* 48, 839–862.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J.J. Ohala & J.J. Jaeger (Eds.) *Experimental Phonology* (pp. 13–44). New–York: Academic Press,

Lindblom, B. (1990). On the notion of possible speech sound. *J. Phonetics,* 18, 135–152.

Locke, J. L. (1983). *Phonological Acquisition and Change*. Academic, New York

Lœvenbruck, H., & Perrier, P. (1997). Motor control information recovering from the dynamics with the EP hypothesis. *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes – Greece, 4, 2035–2038.

MacNeilage, P. F., & Davis, B. L. (2000). On the origins of internal structure of word forms. *Science*, 288, 527–531.

Massaro, D.W. (1987). *Speech perception by ear and eye: a paradigm for psychological inquiry*. London: Laurence Erlbaum Associates.

Mattingly, I.G. (1990). The global character of phonetic gesture. *J Phonetics*. 18, 445–452.

Moon, S.-J., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *J. Acoust. Soc. Am.,* 96, 40-55.

Moore, B.C.J. (2003). *An Introduction to the Psychology of Hearing*, 5th Ed. Academic Press, San Diego.

Nearey, T.M. (1997). Speech perception as pattern recognition. *J. Acoust. Soc. Am.,* 101, 3241-3254.

Nelson, W.L. (1983). Physical principles for economies of skilled movements. *Biol Cybern* 46, 135–147.

Ohala, J. J. 1989. Sound change is drawn from a pool of synchronic variation. In L. E. Breivik & E. H. Jahr (eds.) *Language Change: Contributions to the study of its causes* (pp. 173-198). Berlin: Mouton de Gruyter.

Perrier, P., Loevenbruck, H., & Payan, Y. (1996). Control of tongue movements in speech: the Equilibrium Point Hypothesis perspective. *J. Phonetics*, 24, 53-75.

Pitermann, M. (2000). Effect of speaking rate and contrastive stress on formant dynamics and vowel perception. *J. Acoust. Soc. Am.,* 107, 3425-3437.

Repp, B. H. (1983). Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. *Speech Communication*, 2, 341-361.

Rochet-Capellan, A., & Schwartz, J.L. (2007). An articulatory basis for the labial-to-coronal effect: /pata/ seems a more stable articulatory pattern than /tapa/. *J. Acoust. Soc. Am.*, 121, 3740-3754.

Sato, M., Baciu, M., Lœvenbruck, H., Schwartz, J-L., Cathiard, M-A., Segebarth, C. & Abry, C. (2004). Multistable perception of speech forms in working memory: An fMRI study of the verbal transformation effect. *NeuroImage*, 23, 1143-1151.

Sato, M., Vallée, N., Schwartz, J.L., & Rousset, I. (2007). A perceptual correlate of the Labial-Coronal Effect. *J. Speech Language and Hearing Research* (in press).

Savariaux, C., Perrier, P., & Orliaguet, J.-P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: A study of the control space in speech production. *J. Acoust. Soc. Am.*, 98, 2428–2442.

Schwartz, J.L., Abry, C., Boë, L.J., & Cathiard, M.-A. (2002). Phonology in a Theory of Perception-for-Action-Control. In J. Durand & B. Laks (Eds.) *Phonetics, Phonology, and Cognition* (pp. 254–280). Oxford: Oxford University Press,.

Schwartz, J.L., Abry, C., Boë, L.J., Ménard, L., & Vallée, N. (2005). Asymmetries in vowel perception, in the context of the Dispersion-Focalisation Theory. *Speech Communication*, 45, 425-434.

Schwartz, J.L., Boë, L.J., & Abry, C. (2007). Linking the Dispersion-Focalization Theory (DFT) and the Maximum Utilization of the Available Distinctive Features (MUAF) principle in a

Perception-for-Action-Control Theory (PACT). In M.J. Solé, P. Beddor & M. Ohala (eds.) *Experimental Approaches to Phonology*. Oxford University Press (to appear).

Schwartz, J.L., Boë, L.J., Vallée, N., & Abry , C. (1997). The dispersion-focalization theory of vowel systems. *J. Phonetics*, 25, 255-286.

Serkhane, J.E., Schwartz, J.L., Boë, L.J., & Bessière, P. (2005). Building a talking baby robot: A contribution to the study of speech acquisition and evolution. *Interaction Studies*, 6, 253-286.

Serkhane, J.E., Schwartz, J.L., Boë, L.J., Davis, B.L., & Matyear, C.L. (2007). Infants' vocalizations analyzed with an articulatory model: A preliminary report. *J. Phonetics,* in press.

Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. Davis Jr. & P. B .Denes (Eds.), *Human Communication: A Unified View* (pp. 51–66). New-York: Mc Graw-Hill,

Stevens, K.N. (1989). On the quantal nature of speech. *J. Phonetics*, 17, 3–45.

Stevens, K.N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.,* 111, 1872-1891.

Stevens, K.N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In Wathem-Dunn, W. (Ed), *Models for the Perception of Speech and Visual Form. Cambridge, MA: MIT Press.*

Vilain, A., Abry, C., Badin, P., & Brosda, S. (1999). From idiosyncratic pure frame to variegated babbling: Evidence from articulatory modelling. In *Proceedings of the 14th International Congress of Phonetic Sciences*, San Fransisco, pp. 2497–2500.

Viviani, P., & Stucchi, N. (1992). Biological movements look uniform: evidence of motor-perceptual interactions. *Journal of Experimental Psychology: Human Perception and Performance,* 18, 603–623.

Warren, M.R (1961). Illusory changes of distinct speech upon repetition – The verbal transformation effect. *British Journal of Psychology*, 52, 249-258.

Whalen, D. H., Benson, R. R., Richardson, M.L, Swainson, B., Clark, V. P., Lai, S., Mencl, W. E., Fulbright, R. K., Constable, R. T., & Liberman, A. M. (2006). Differentiation of speech and nonspeech processing within primary auditory cortex. *J. Acoust. Soc. Am,* 119, 575–581.

Whalen, D. H., & Liberman, A.M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, 237, 169–171.

Wood, S. (1979). A radiographic analysis of constriction locations for vowels. *J. Phonetics*,7, 25–43.

Watkins, K.E., Strafella, A.P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia,* 41, 989–994.

Auditory theories
(e.g. Nearey, Massaro)

no-link, no gap

Motor theories
(e.g. Liberman, Fowler)

full-link, no gap

Theories of the perceptuo-motor gap

Ohala's theory of the sound change: *The gap drives sound changes*

Lindblom's adaptive variability theories: *The gap leads to system optimization*

Stevens' Quantal Theory: *The gap shapes categories*

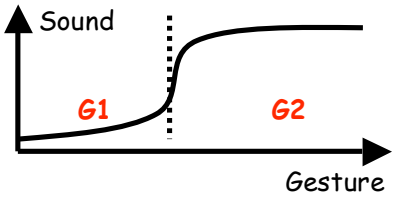

Figure 1: The perceptuo-motor gap in classical speech communication theories

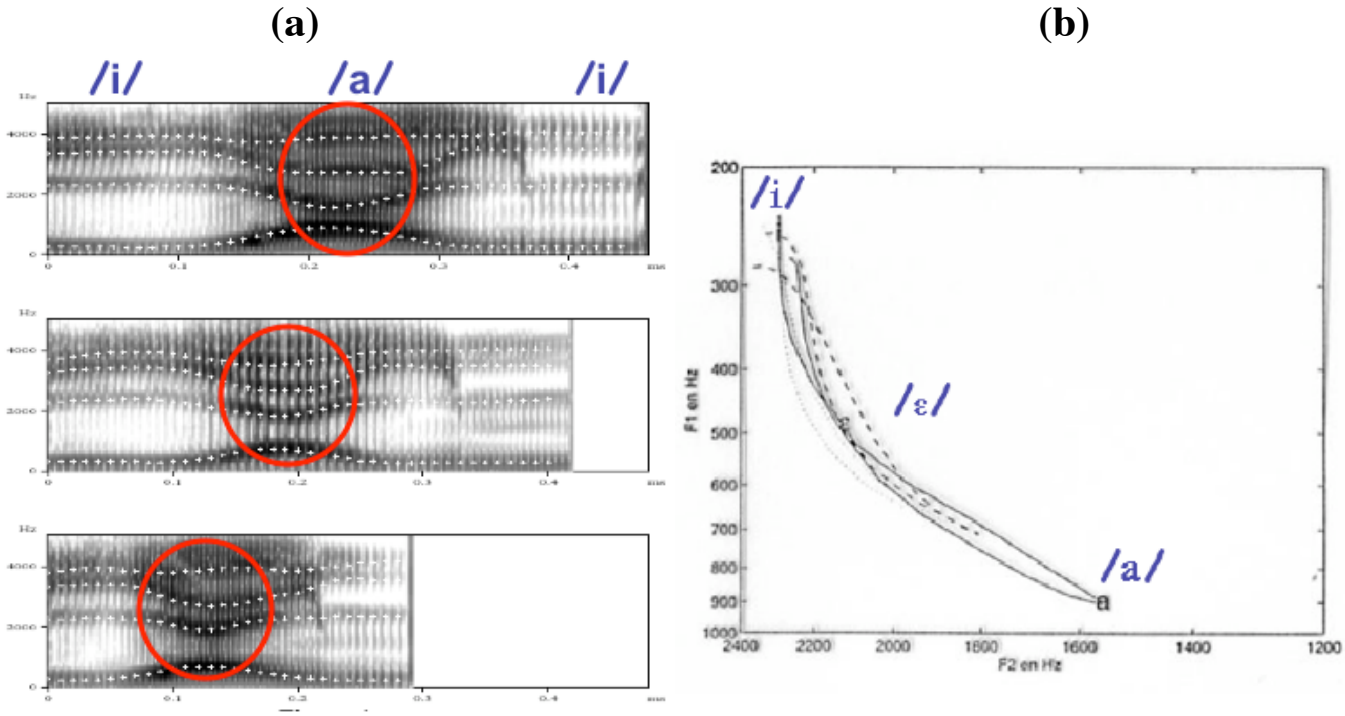

Figure 2: A typical case of vowel reduction
(a): Spectrograms of [iai] utterances in French, with decreasing strength of articulation and increasing speed from top to bottom.
(b): Corresponding formant trajectories in a (F1, F2) space.
From Loevenbruck et al., 1997

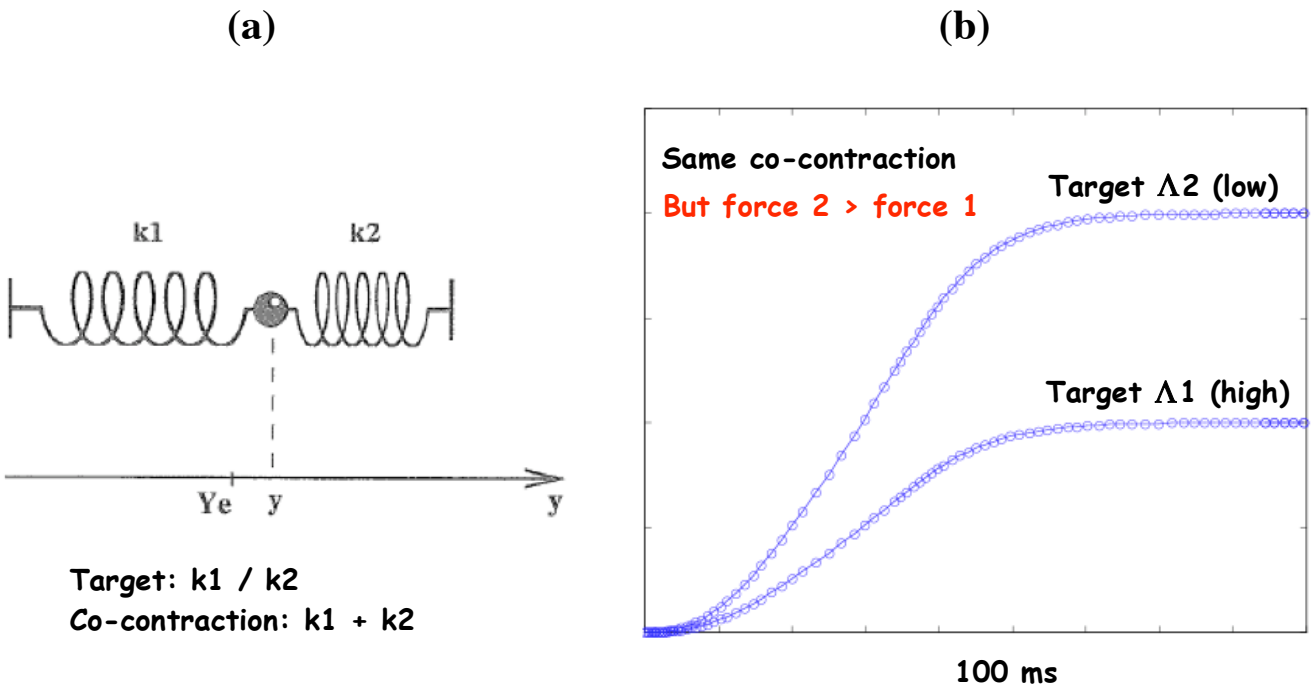

Figure 3: Simulating trajectories from a closed position to a near high target Λ1 or a far low target Λ2 in the framework of second order agonist-antagonist models. (a): the model; (b): simulated trajectories, for equal values of the cocontraction, but different forces applied.

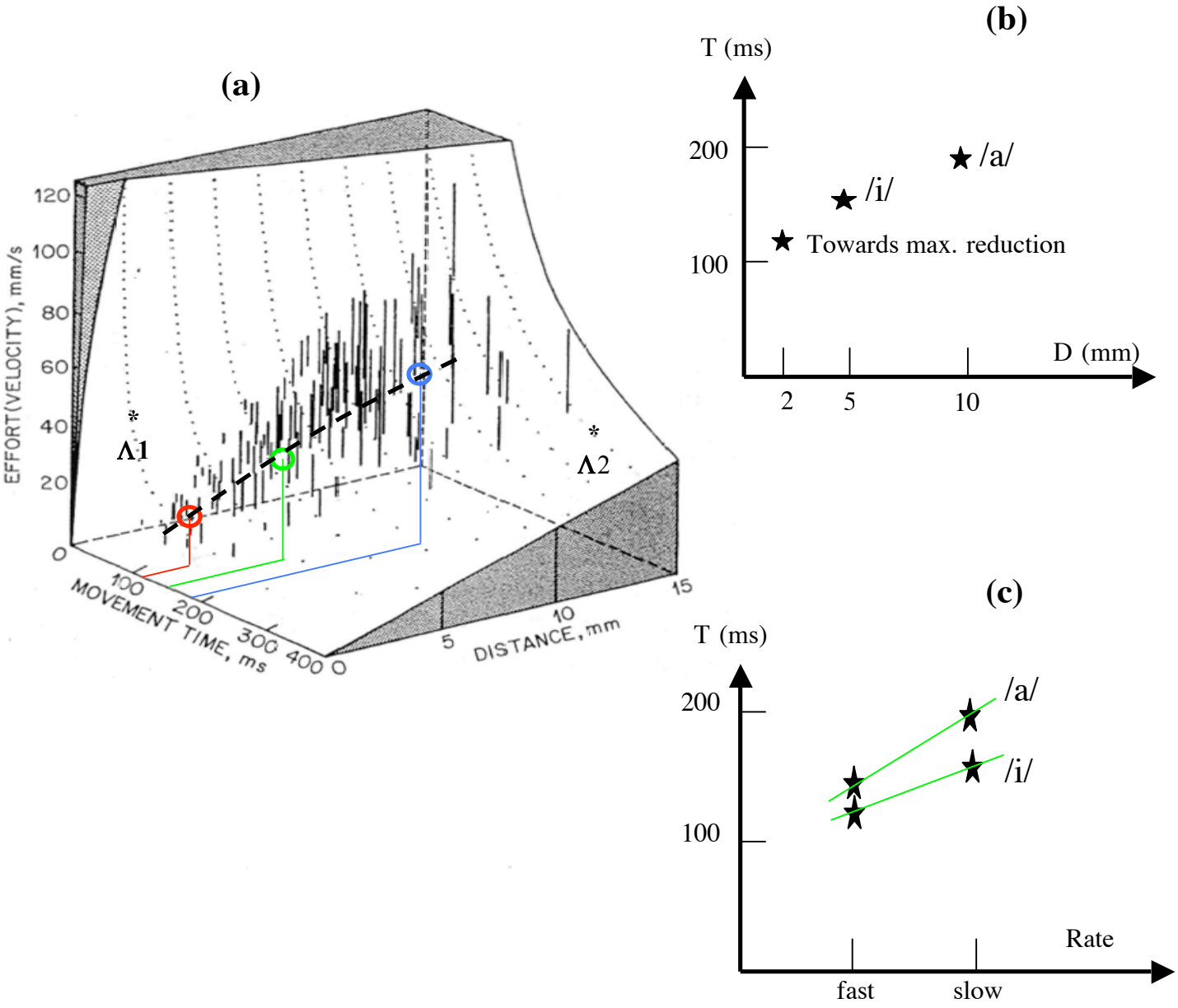

Figure 4: Relationships between effort E, distance D and movement time T in second order gestures
(a): Theoretical predictions about the (E,D,T) relationships (concave dotted lines) together with real data from jaw stroke analyses (vertical lines) (from Nelson, 1983). Λ1 and Λ2 are respectively a near and a far target. The thick dotted line displays the (E,D,T) link for realistic jaw movements
(b) Likely relationships between T and D for realistic jaw movements
(c) Derived predictions about movement time for far (e.g. [a]) and close (e.g. [i]) targets, as a function of speech rate

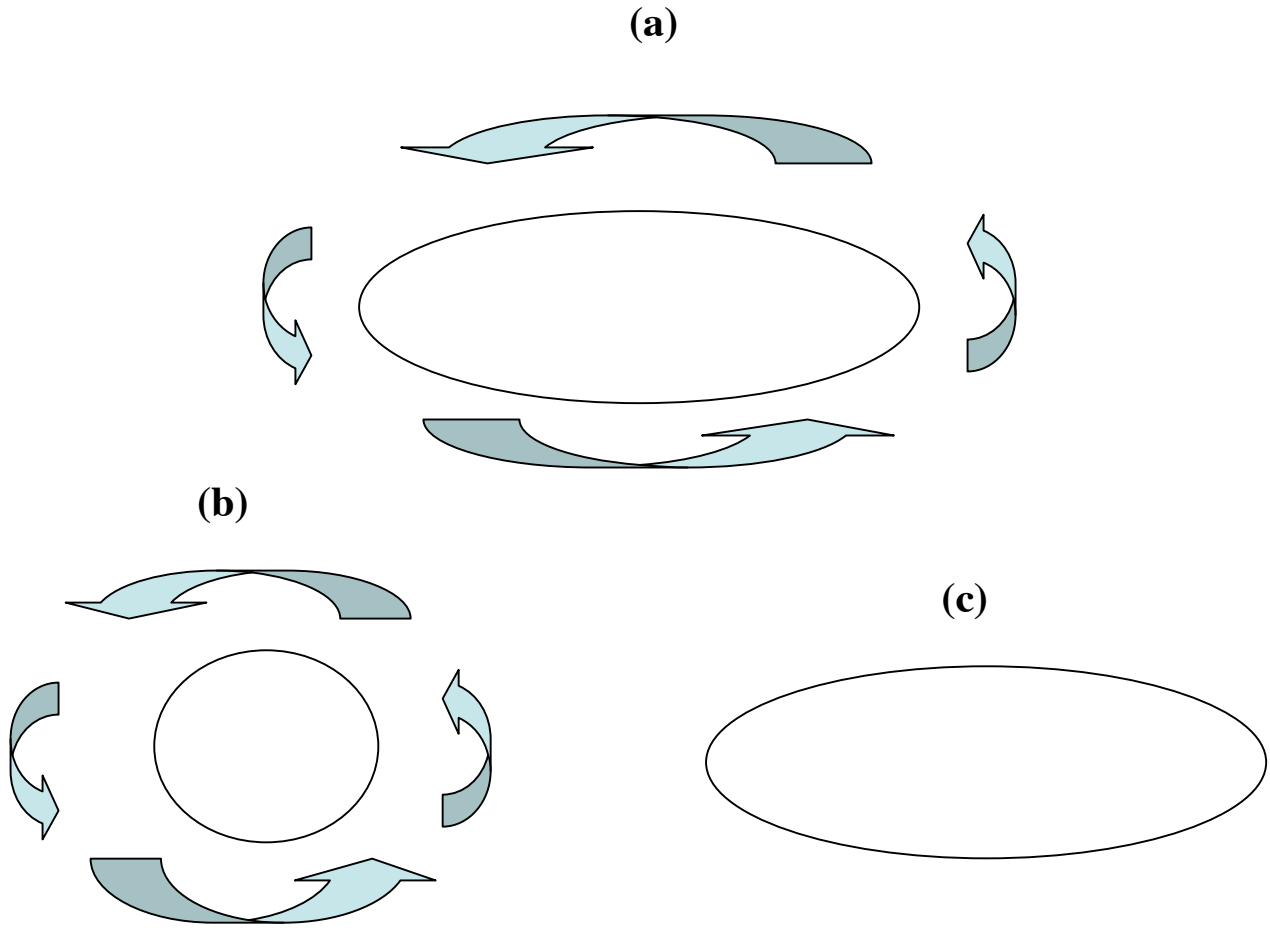

Figure 5: The " $v = r^{1/3}$ " law of human gestures
and the role of procedural knowledge in action perception
(a): to be perceived as stable in speed, a human movement must obey the " $v = r^{1/3}$ law, with quicker movements in straight portions
(b): applying an "elliptic" speed pattern along a circular shape produces the perception of an elliptic shape (c)

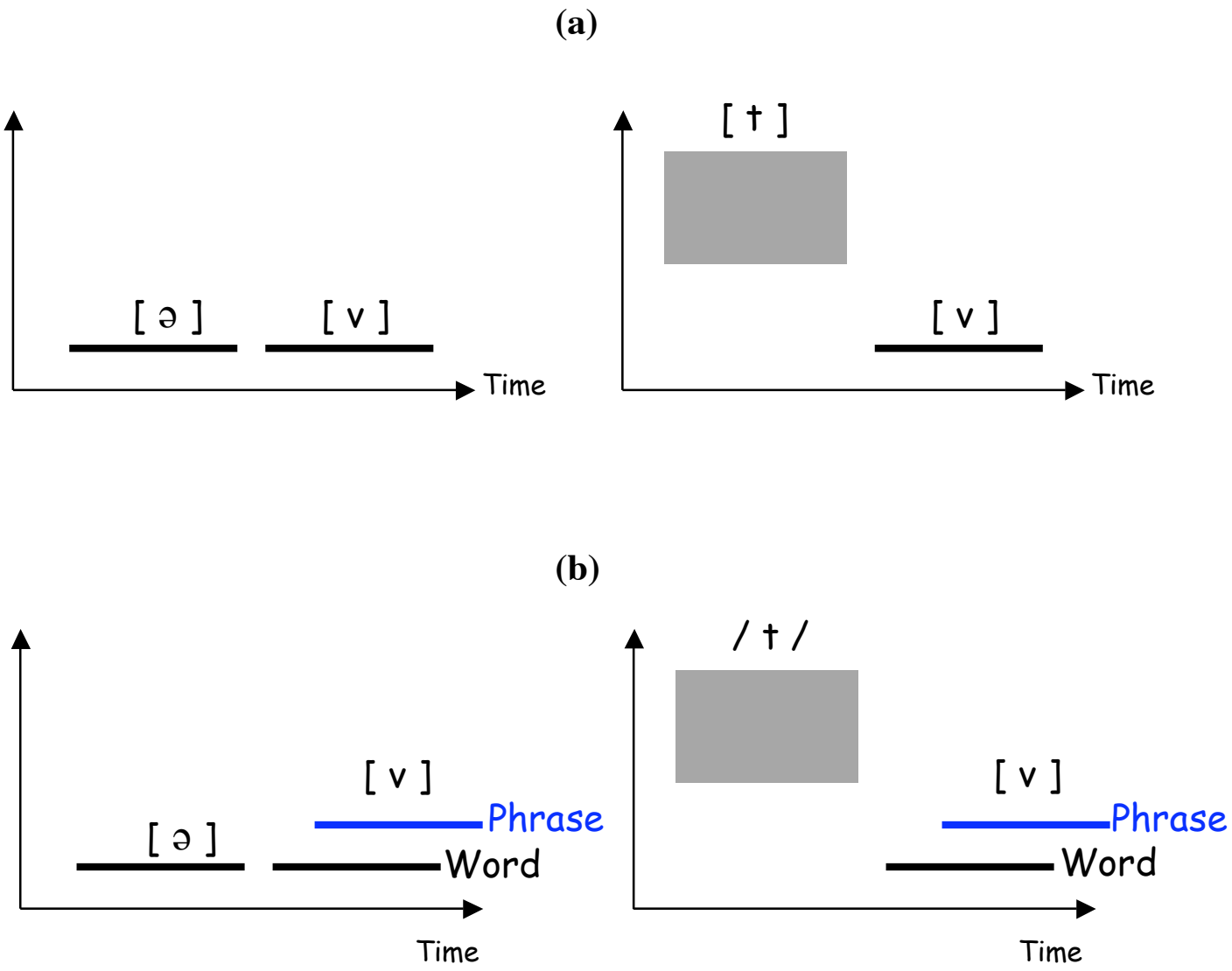

Figure 6: Forward masking predictions
of the assimilation compensation results by Kuzla et al.
(a): forward masking decreases the audibility of the [v] voicing bar,
more with an [ə] context (left) than with a [t] context (right)
(b): forward masking should decrease the audibility of the [v] voicing bar, more
with a word-boundary than with a phrase-boundary context